SC17 StarLight/iCAIR/MREN Networking Demonstrations

Joel Mambretti
Director of the International Center for Advanced Internet Research,
Northwestern University
Summer 2018

With their multiple national and international research partners, the International Center for Internet Research at Northwestern University (iCAIR); the StarLight consortium, which manages the StarLight International/National Communications Exchange Facility in Chicago; and the Metropolitan Research and Education Network (MREN) showcased demonstrations of next generation networking for data intensive science at the SC17 International Conference for High Performance Computing, Networking, Storage and Analysis in Denver, Colorado (November 11-17, 2017). These demonstrations were supported by the SC17 SCinet engineers who implement and manage the conference network and the SCinet WAN.

Below is a list of the advanced technologies that were showcased:

Services of large scale data intensive science
Global Research Platform
Highly distributed science DMZs
Software Defined Networks
Software Defined Exchanges
Network Slicing/Stitching
MultiDomain East West L2 Protocol - NSI 2.0
AutoGOLE Dynamic Path Exchanges
Large Scale Airline Data Transport Over 100 Gbps WANs
Science Clouds, Including the Open Science Data Cloud
100 Gbps Radio Astronomy Networking
100 Gbps Weather Services Networking
Multi-100 Gbps WAN Services Developed by the NASA Goddard Space Flight Center
Software Defined WAN Services
100 Gbps Data Transfer Nodes (DTN)
SCinet DTN as a Service
Advanced Large Hadron Collider Networking Services
Network Measurement and Analytics at 100 Gbps
Big Data Express Services
mdtmFTP High Performance WAN Transport Protocol
Pacific Research Platform/Prototype National Research Platform
Kubernetes for Large Scale Science

SAGE 2 Scientific Visualization
European Space Agency's (ESA) Copernicus Earth Observation
Open Control Waves

The focal data intensive petascale sciences for these demonstrations included high energy physics, astrophysics, radio and optical astronomy, bioinformatics, geophysics, oceanography, atmospheric modeling, weather prediction, climate modeling, space exploration, aeronautic data analysis, data science, and scientific visualization. Traditional networks cannot support the large scale global data flows required by these sciences. Consequently, a network research consortium has established multiple collaborations to design, develop, implement, and showcase next generation network architecture capabilities that can provide the services required by these sciences.

The components of the services that were showcased included those based on highly distributed science DMZs, for example, distributed environments such as the Global Research Platform (GRP). A major component of this platform is Software Defined Networking (SDN), including capabilities for steering data flows with the Jupyter science workflow tools, which can be integrated as orchestrators with network resource provisioning tools. A related orchestration mechanism that was demonstrated was the P4 network programming language, which can be customized to steer international WAN high performance data transfer services at 100 Gbps. The SDN architecture per se is single domain oriented. To provide multiple domain capabilities, this research consortium is developing and demonstrating services, architecture, and technologies for Software Defined Exchanges (SDXs) that can directly control dynamic 100 Gbps paths. A key SDX is the StarLight NSF International Research Connections (IRNC) SDX. These capabilities are being demonstrated in partnership with National Center for High-performance Computing (NCHC), in TaiwanKorea Institute of Science and Technology Information (KISTI) in South Korea), Compute Canada, the University of Warsaw and Poznan Supercomputing and Networking Center (PSNC) in Poland, the University of Sydney and Australia's Academic and Research Network (Australia), the European Organization for Nuclear Research (CERN/CERNLight) in Switzerland, the Singapore high performance computing center and other research sites.

Other SDX capabilities that were demonstrated at SC17 included advanced switching capabilities, showcasing multiple switching architectures and technologies, including highly programmable sliceable SDN switches. Several SC17 IRNC SDX demonstrations included a fundamentally new concept of "consistent network operations," where stable load balanced high throughput workflows crossing optimally chosen network paths, preset demarks to accommodate other traffic, provided by autonomous site-resident services, dynamically interacting with network-resident services, in response to requests from the science programs' principal data distribution and management systems.

Components of SDXs included the Network Service Interface (NSI 2.0), a standard east west L2 multiple domain East West protocol developed by the Open Grid Forum. These SC17 SDX demonstrations also showcased the capabilities of the MEICAN provisioning tools, created by RNP, Brazil's national R&E network, which provide an API for NSI.

Another SDX component that was showcased was the AutoGOLE switching service of the Global Lambda Integrated Facility (GLIF). One part of these demonstrations was a prototype international large scale high performance transport service point-to-point (P2P) E2E LHC WAN service for large capacity data streams based on AutpGOLEs and NSI. Another related SDX supported demonstration showcased a prototype service for transporting large scale airline data, a joint project with SURFnet, the University of Amsterdam, NetherLight, iCAIR and KLM-Air France.

Another series of demonstrations showcased network edge integrations with computational resources, including science clouds, such as the Open Science Data Cloud, machine learning facilities, computer science testbeds, such as the Global Environment for Network Innovation (GENI) and the Chameleon Could testbed, supercomputing centers, and multiple supercomputing centers interlinked with 100 Gbps channels that can support mobile, scalable processes.

Another set of demonstrations showcased prototypes services for radio astronomy, which includes support for existing instruments and data stores, such as the Sloan Digital Sky Survey (a collaboration with Johns Hopkins University), and planning for new instruments such as the LSST and SKA (a collaboration with the University of Sydney and AARnet).

Other sets of science based SC17 demonstrations showcased a new method for supporting weather science workflows, including international WAN high performance data transfers. This project is a joint partnership of iCAIR, NCHC in Taiwan, the National Oceanographic and Atmospheric Administration, and the National Center for Atmospheric Research.

Another set of geophysical, weather, atmosphere science and space exploration related demonstrations supported by the StarLight facility have been staged by the NASA Goddard Space Flight Center's High-End Computing (HEC) group in Greenbelt Maryland. This group has been addressing the requirements of big data-based space exploration and geophysical science, including weather prediction, atmospheric research, climate modeling and oceanography. At SC16, the HEC networking group demonstrated 200 Gps disk to disk data transfers over WANs. At SC17, they demonstrated 400 Gbps disk to disk over a national WAN.

Also, showcased through SC17 demonstrations were globally distributed Data Transfer Nodes (DTNs) interconnected with 100 Gbps paths. DTNs are network appliances optimized for transporting data at 100 Gbps over SD-WANs, including large capacity single streams. These demonstrations included showcasing capabilities of specialized 100 Gbps DTNs interconnected globally over many thousands of miles, and innovative DTN middleware optimized for large scale

science workflows. They were also optimized for sending large capacity streams with high performance disk to disk across WANs.

The iCAIR PetaTrans 100 Gbps Data Transfer Node (DTN) demonstrations were directed at showcasing large scale WAN services for high performance, long duration, large capacity single data flows. iCAIR is designing, developing, and experimenting with multiple designs and configurations for 100 Gbps Data Transfer Nodes (DTNs) over 100 Gbps Wide Area Networks (WANs), especially trans-oceanic WANs, PetaTrans – high performance transport for petascale science, showcased at demonstrations at SC17.

A related set of demonstrations was a collaboration with SCinet that has implemented a DTN in SCinet, Because the SC research exhibitions are increasingly using multiple 100Gs in their booths, it was important for them to have access to DTN services. Having a DTN in SCinet enabled them to focus on their demonstrations, and reduced the need to spend time configuring networks and systems, This SC17 DTN provided a) a network and system test point (software stack, architecture, connectivity and performance, including over WANs) for large data transfer projects before and during the conference, b) a platform for large flow projects before and during the conference c) additional capacity for large flow projects before and during the conference d) access to experimental scalable DTN technology, including over WANs.

At SC17, the StarLight facility supported an ATLAS community demonstration of new network based distributed storage techniques for ATLAS experiments, new techniques for optimizing ATLAS workflows using machine learning and real time analytics, and new techniques for optimizing data flows using machine learning. The ATLAS community developing machine learning based simulation and analysis techniques on a distributed platform comprised of shared GPU machines connected by high performance WANs networks to enable transparent streaming of data using a distributed Ceph file system (OSiRIS).

Another series of demonstrations showcased granulated measurements and analytics at 100 Gbps, which can be used to discover patterns invisible to traditional network measurement tools and which can be used to automate network operations that today are addressed through manual processes.

Other StarLight SDX supported measurement demonstrations were based on a partnership with the NSF IRNC Antarctic Infrastructure Modernization for Science (AIMS) project, which is being led by the University of Massachusetts. Network measurement and monitoring are instrumental to network operations, planning, and troubleshooting. However, increasing line rates (100+Gbps), changing measurement targets and metrics, privacy concerns, and policy differences across multiple R&E network domains have introduced tremendous challenges in operating such high-speed heterogeneous networks, understanding the traffic patterns, providing for resource

optimization, and locating and resolving network issues. The AIMS project is addressing these issues.

Also, the NSF IRNC StarLight SDX supported the FNAL demonstrations of the Big Data Express (BDE) project, which has released several software packages, including the BDE WebPortal. This FNAL developed web portal allows users to access BigData Express data transfer services. Another software stack is a BDE scheduler, which schedules and orchestrates resources at BDE sites to support high-performance data transfer. A third is BDE AmoebaNet, which is a SDN-enabled network service that enables "application-aware" networks. It allows applications to program network at run-time for optimum performance. These software packages can be deployed to support three types of data transfer: real-time data transfer, deadline-bound data transfer, and best-effort data transfer. For the SC17 FNAL demonstrations, BDE software were used to demonstrate bulk data movement over wide area networks. One goal was to demonstrate that BDE can successfully address the high-performance and time-constraint challenges of data transfer to support extreme-scale science.

The StarLight SDX also supported demonstrations of Fermilab network research that newly developed high-performance data transfer tool, called mdtmFTP, which maximize data transfer performance on multicore platforms. mdtmFTP has several advanced features. First, mdtmFTP adopts a pipelined I/O design. A data transfer task is carried out in a pipelined manner across multiple cores. Dedicated I/O threads are spawned to perform I/O operations in parallel. Second, mdtmFTP uses a particularly designed multicore-aware data transfer middleware (MDTM) to schedule cores for its threads, which optimize use of underlying multicore core system. Third, mdtmFTP implements a large virtual file mechanism to address the lots-of-small-files (LOSF) problem. Finally, mdtmFTP unitizes multiple optimization mechanisms – zero copy, asynchronous I/O, batch processing, and pre-allocated buffer pools – to improve performance. In these demonstration, mdtmFTP is used to optimized bulk data movement over long-distance wide area networks.

Another set of StarLight SDX supported demonstrations expanded large scale research platform projects. To date, Science DMZs have been primarily local to campuses. The Pacific Research Platform (PRP) is a regional Science DMZ interconnecting multiple campus science DMZs. The initial steps have been taken to explore the possibility of a National Research Platform. Currently, the StarLight community has been implementing the building blocks required to create a Global Research Platform (GRP). The initial stages of this platform were demonstrated at SC17.

In partnership with the Pacific Research Platform Project, CENIC, Pacific Wave, and others, iCAIR and StarLight supported demonstrations of software-based route-servers on IRNC Pacific Wave and StarLight devices to showcase a Border Gateway Protocol (BGP) pilot capability. The route-server model creates a BGP-signaled control plane across which participants can announce routes, with the traffic flowing across separate data-plane connections. Among the

goals of the pilot is to explore the scalability of such a model for separating high-performance, data-intensive workflows from general academic and administrative traffic. Each route-server is configured to operate a BGP "view" of PRP/AS395889. AS395889 is the Autonomous System Number (ASN) assigned by the American Registry of Internet Numbers (ARIN) specifically for use by the PRP in response to a request submitted by Pacific Wave.

This project is also exploring the use of the PRP/AS359889 to facilitate interconnection among participating institutions to selectively announce reachability of resources specifically supporting high-performance research, including data placement/distributed storage, HPC/compute virtualization/GPU computing, network function virtualization (NFV), named data networking (NDN)/content-centric networking (CCN), integrated computing network, among other data-intensive efforts, as well as network performance measurement and analysis infrastructure. In forming the control plane at a given location (exchange), participants peer with the router-server using external-BGP (eBGP) with each participant's eBGP speaker configured as a route-server client. A participant can establish a single session with a route-server and receive routes from all the other route-server participants. This model can simplify the tasks involved in designing and implementing BGP policy. The data-plane connections are a combination of manually and dynamically provisioned circuits. Dynamically-provisioned circuits are orchestrated using the RNP-developed MEICAN web user interface as part of the GLIF AutoGOLE/NSI effort. This initial set of route servers is running a BGP daemon from the open-source Quagga routing software suite:

A related demonstration resulted from a collaborative partnership with UCSD that is exploring specialized software stacks that could be useful for creating a multi-institution, hyperconverged science DMZ, using Kubernetes as a core orchestrating resource. At SC17, this consortium demonstrated the utility of the services provided by a prototype model to support data intensive science. The JupyterHub Kubernetes Spawner enables JupyterHub to spawn single-user notebook servers on a Kubernetes cluster. JupyterHub can be run from inside Kubernetes. Consequently, many JupyterHub deployments can be run with Kubernetes only, eliminating a need for using scripts such as Ansible, Puppet and Bash. It has utilities for integrated monitoring and failover for the hub process.

In partnership with the University of Illinois (UIC) at Chicago's Electronic Visualization Laboratory (EVL), iCAIR and StarLight supported a prototype cloud-based SAGE2 scientific visualization service that deploys Docker containers running on a SAGE2 server at UIC in Chicago. The SAGE2 browser was connected to high-resolution display monitors on the SC17 showfloor. SDN networks were used to transfer information between the two locations. The SAGE2 browser accesses large files and collections of many small files at StarLight. SDN/SDX was used to find and create end-to-end paths across WANs on which to send the files. The StarLight SDX coordinated the routes to select best paths across the 100 Gbps WANs.

Another set of SC17 demonstrations showcased an international collaboration that provides capabilities for sharing satellite repository data. The National Computational Infrastructure (NCI) facility at the Australian National University in Canberra is hosting a major regional data repository as part of the European Space Agency's (ESA) Copernicus Earth observation satellite project. Copernicus, comprised of six Sentinel satellites, provides data that assists policymakers in addressing environmental policies and reacting to emergencies, including natural disasters or humanitarian crises. The regional hub is Copernicus' master data repository for the South-East Asia and South Pacific region, with NCI's high-performance compute capabilities to be used to crunch data from the satellite program. Satellite data is collected through the ESA's IntHub landing station, and the transferred to Australia across two separate redundant intercontinental data paths through international R&E networks and exchange points. Transferring ESA Copernicus satellite data from Europe to Australia presents significant network stack tuning requirements to optimize data transfers across the extended delay product that Australia's geographic position presents. The SC17 demonstrations showcased a variety of methods illustrating optimal workflows for reliable and timely transfer of data products from the ESA Hub in Athens, Greece hosted by the Greek Research and Technology Network (GRnet). Direct data transfer from Athens to Canberra, cascaded Squid proxies and staged data transfers used a number of file transfer tools including Aspera, gridFTP and mdtmFTP. These demonstrations were staged by ICM at the University of Warsaw in Poland, ACRC in Singapore, and StarLight and iCAIR in the US.

Another set of SC17 demonstrations, staged by the NRL, Caltech and iCAIR, showcased optical disaggregation through Open Control Waves (OCW) based on 1 Tbps set of waves from the StarLight booth to the CalTech booth, both anchored/terminated by DCIs. In these demonstrations, individual wavelength paths on and off based on use case scenarios as an OCW API were used to provision dynamically an additional 100G network path from one booth to another using the SCinet DTN system. This could be considered a type of a network cloud-bursting equivalent - or a large-scale surge bandwidth use case.

In preparation for SC17, a project was initiated to evaluate the standards, technologies, and switches that are emerging for support for 400 Gbps Ethernet NICs. This project has led to the initial plans for a showcase of these 400 Gbps NIC capabilities for SC18.

"Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Networking and Information Technology Research and Development Program (NITRD). Any mention of specific commercial products or reference to commercial organizations is for information only; it does not imply recommendation or endorsement by NITRD."

The Networking and Information Technology Research and Development (NITRD) Program

Mailing Address: NCO/NITRD, 2415 Eisenhower Avenue, Alexandria, VA 22314

Physical Address: 490 L'Enfant Plaza SW, Suite 8001, Washington, DC 20024, USA Tel: 202-459-9674, Fax: 202-459-9673, Email: nco@nitrd.gov, Website: https://www.nitrd.gov